ASPECT : Audio SPatial Environment for CommunicaTion - as a Three Dimensional Auditory Interaction Tool.

Tatsuhiro YONEKURA, Nobuyuki ARIYOSHI and Yoshiki WATANABE

Department of Computer and Information Science, Faculty of Engineering

Ibaraki University

4-12-1 Naka-narusawa cho, Hitachi, Ibaraki 316 JAPAN

Abstract - This paper introduces a new methodology of human interface through the three dimensional virtual space created in the computer by using the audio signal. That is, the user can access three dimensional (3D hereinafter) virtual space by her/his vocal and auditory ability. As an input function the system contains a multiple three dimensional locator which locates several 3D pointers in the virtual space, and as an output function a certain type of 3D sound generator. By showing some of the demonstrations of the interface system, the effectiveness in a sense of *direct manipulation* as an spatial interaction tool can be confirmed.

key words- **ASPECT, sound source localization, 3D locator, 3D interaction tool.**

1.Introduction

In the last few years multi-media interface technology is widely improved because a variety of interactive device becomes simultaneously and effectively used, especially in the field of 3D spatial interaction such as Data Glove with 3D graphics on Head mounted display(1)(2). These devices have realized high level of *direct manipulation* so that they are able to be utilized for the field of so-called 'Virtual Reality'. From the viewpoint of popularization , however, these devices have more or less disadvantages because of their high cost and heavy computing load.

In addition, our vocal ability and auditory ability are closely tied with each other for communication scheme so that the audio media must be used for more comfortable and smooth interaction with computers.

In this paper we propose a new 3D interaction tool of comparatively simple mechanism by *sound source localization* with lower costs and lighter computing loads which uses audio information from/to the computer.

That is, the system computes the 3D location of the user in the virtual space from the sound that she/he makes, and the user acknowledges where she/he is by the 3D graphics and/or the 3D audio signal that the system makes. In this sense let us call this system **ASPECT** for Audio SPatial Environment for CommunicaTion. We first give the overview idea and the high level functionality of this interface system in the next Section. Then we explain the concrete implementation of it in two detailed functions the subject handling and motion acknowledgment in Section 3. In Section 4 the experimental results are demonstrated about the prototype of our ASPECT system.

2. Overview

Fig.1 shows overview of the basic ASPECT system. Four point microphones set, AD converter, 3D sound locator, 3D audio signal generator (these two may be realized by the software in a computer system), DA converter and headphones set (or the speaker system) are the major elements of the system. Hereinafter we use the term *subject* as the manipulating subject controllable by the user in the virtual space and the term *object* as the object to be

0-7803-1363-1/93 $3.00 © 1993 IEEE

accessed by the subject in the same space. Note that both subject and object are located at a certain points which have Cartesian coordinates (x,y,z). The requirements for the system are :
a). the user can move the subjects via the 3D sound locator triggered by her/his vocal ability (**multiple 3D subject handling function**)
b). the user can acknowledge the relative position from the associated subject to the object via 3D audio signal generated by the system provided that she/he has ordinary auditory ability (**3D audio function**)
c). satisfying the above two requirements the visual information via the graphics of CRT display is not necessarily involved in the system for performance of 3D interaction (**real-time loopback interaction**)

To satisfy a) above, we apply four point microphones set as Fig.2 to obtain the difference in sound pressure which gives information about the direction of sound source. The algorithm is explained later in more detail. And to satisfy b), we utilize the 3D audio signal characterized by the difference in sound pressure.

Thus the goal of the ASPECT is to support spatial man-machine interaction environment by using the complementary audio media, or more strongly, to realize the 3D man-machine interface environment without the visual information in the presence of audio information which even visually disabled people can fairly work with.

3.Implementation

Suppose there are N controllable subjects and N accessible objects in the system. We let them denote s_1 through s_N and o_1 through o_N respectively. In order to make correspondence between subjects and objects, assume that n'th subject is corresponding to the n'th object (more than two objects can be identical ($o_l=o_m$) , so one to one correspondence is not necessary between the subjects and objects). Let location of n'th subject and object be (sx_n,sy_n,sz_n) and (ox_n,oy_n,oz_n) where n=1,2,.,N. Our problems are
1) how to manipulate each subject separately effectively
2) how to let the user acknowledge where the corresponding object is from her/his associated subject location
So, in the following Subsections we describe, as function of 3D sound locator, the mechanism of handling subjects by 3D sound localization and then, the clue of making 3D sound by the 3D sound generator to let the user acknowledge where the object is.

3.1 Handling 3D subjects

Our method of handling 3D subjects is summarized as follows.

To support the multiple subject handling function using the single microphone system above, multiple sound source method is applied. That is, several phonemes (e.g, 'a', 'i', 'o',etc. corresponds to the first, second and third subject) are sampled before execution and do the matching between the sample phoneme and input phoneme to identify the sound source. The subject corresponding to the matched phoneme is given with the appropriate motion according to the difference in sound pressure sensed among the four microphones. So the user can let a particular subject move by her/his vocal sound as she/he controls.

In designing our prototype system, correlation coefficient between the sample phonemes and the input phoneme in the frequency domain is applied as the goodness of matching. The determination of matching is done when both of the two conditions are satisfied :(1) goodness of matching is greater than a certain threshold (e.g, 0.7) and (2) goodness of matching marks the highest score among N samples.

Only the difference in sound pressure among the microphones is used for manipulation scheme in the prototype. In Fig.2 we assume the four microphones M1 through M4 of the set placed at the point (-1/2,-1/2,-1/2), (1/2,-1/2,-1/2), (-1/2,1/2,-1/2) and (-1/2,-1/2,1/2) in the coordinate system of the virtual space and the 3D vectors to them are called $\mathbf{m}_1$ through $\mathbf{m}_4$. Thus the unit vectors of each component (1,0,0), (0,1,0) and (0,0,1) are represented by ($\mathbf{m}_2$-$\mathbf{m}_1$), ($\mathbf{m}_3$-$\mathbf{m}_1$) and ($\mathbf{m}_4$-$\mathbf{m}_1$) respectively, and the origin of the coordinates (0,0,0) is represented by vector sum of the three unit vectors above plus vector $2\mathbf{m}_1$, which is ($\mathbf{m}_2$+$\mathbf{m}_3$+$\mathbf{m}_4$-$\mathbf{m}_1$).

Now suppose we have sound signals $s_1(t)$, $s_2(t)$, $s_3(t)$ and $s_4(t)$ input into the microphone M_1 through M_4 from $t=t_0$ to t_0+Kd_o where d_o is the sampling period. Then the power of sound input into i'th microphone p_i becomes

p_i= square root $(S_k |s_i(t_0+kd_o)|^2)$ (k=1,2,..,K ; i=1,2,3,4)

At this moment the sound source point is supposed to locate at **v**=(x,y,z) by physical laws where **v** satisfies the following equations.

$p_1|\mathbf{v}-\mathbf{m}_1|^2 = p_2|\mathbf{v}-\mathbf{m}_2|^2 = p_3|\mathbf{v}-\mathbf{m}_3|^2 = p_4|\mathbf{v}-\mathbf{m}_4|^2$

sign(x) = sign(p_2 - p_1)
sign(y) = sign(p_3 - p_1)
sign(z) = sign(p_4 - p_1)

Provided that p_1, p_2, p_3 and p_4 are comparable the rough value of (x,y,z) is fairly estimated by the following equation.

x = p_2 - p_1
y = p_3 - p_1
z = p_4 - p_1

Hereinafter we use this simple sound source localization algorithm in the prototype.

So our next step is how to decide an amount of motion given to a subject. The easiest way is to use the vector v as the velocity of a subject as follows. Suppose the n'th subject is selected by the matching scheme above;

$sx_n(t+Kd_o) = sx_n(t) + ex$
$sy_n(t+Kd_o) = sy_n(t) + ey$
$sz_n(t+Kd_o) = sz_n(t) + ez$

gives the magnitude and the direction of motion of the subject. It may be more effective way of look and feel to add the moment terms as $b(sx_n(t) - sx_n(t-Kd_o))$ and so on, to each equation.

3.2 Three dimensional sound generation

When the position (sx,sy,sz) and velocity (vx,vy,vz) are given to a subject, the system should output an appropriate feedback signal to the user for real-time acknowledgement. The general way is to generate the audio signal from the direction ((ox-sx),(oy-sy),(oz-sz)) to be given to the user. In case of multiple subject multiple object system it is satisfying to reply with the sound from the direction of corresponding object to the associated subject.

In this case while the user pronounces the n'th phoneme the system outputs the sound from n'th object $((ox_n-sx_n),(oy_n-sy_n),(oz_n-sz_n))$ so that the user can tell the direction and the distance to the target object. It is able for the user to change the manipulating subject to another one (say (n+1)'th) by pronouncing (n+1)'th phoneme while the system replies with the echo back from (n+1)'th object $((ox_{n+1}-sx_{n+1}), (oy_{n+1}-sy_{n+1}),(oz_{n+1}-sz_{n+1}))$.

Although there are various way to generate the 3D sound [3][4] reported so far, we apply in the prototype the simplest stereo method with difference in volume between left and right speaker of a pair of headphones. We use four types of sound for distinction among upper-front (for indication to the direction of positive Y with negative Z) , upper-back (for positive Y with positive Z), lower-front (for negative Y with negative Z) and lower-back (for negative Y with positive Z). By combination of these two features, which is difference in volume and these four sound types, we work with the prototype for easy 3D spatial task such as accessing 3D points with multiple subjects without any visual information. We show them in the next Section.

4. Demonstration

Now we show some demonstration of the ASPECT system. In this prototype we use a PC with 32 bit CPU. In the experiment we set three subjects with three objects in the virtual space. The concrete scheme of experiment is as follows.

a). Three objects numbered #1, #2 and #3 are placed somewhere in the virtual space (ox_n,oy_n,oz_n) where n=1,2,3 ; which actually are arranged by the computer using random values of $(-100<ox_n<100, -100<oy_n<100, -100<oz_n<100)$.

b). The user is prompted to input three different phoneme into microphones as sample phonemes (usually vowel sounds e.g, 'a', 'i' and 'o' which are known to be well distinguishable as shown in the Table 1: Correlation among five vowel sounds) while the system performs FFT (Fast Fourier Transform) to these samples and saves the results for preparation of the 'game'. The setup sequence is done and let the ASPECT start. The sampling period is 100 micro seconds as d_0 and K=64 periods is contained in each sample phoneme data so the frequency bandwidth is from 100 Hz to 6.5 kHz.

c). The system initiates real-time loopback interaction between human and machine by combining the subjects handling and echo back function of the ASPECT.

c-1). The user voices again one of the previously set phonemes from appropriate position of the set. The system at this moment computes which subject is to move by matching the phoneme with sample phonemes and what the direction and amount the motion is by the simple sound source localization algorithm in Section 3. The way of sound sampling is the same as done in sample phoneme input above.
If no sample is matched the system discards the input phoneme and try step c-1) again.

c-2). In order to hit the object by the subject the user almost simultaneously listens the direction and sound type from the system via the headphones while she/he voices to the set. The system computes the gap between the two vectors i.e, the associated subject's location and the corresponding object's location. The system generate the appropriate sound type with appropriate amount of the sound for left speaker and right speaker. The shape of each subject is a sphere with a radius of 6. The condition of *reach* is made when an object point touches or passes through the corresponding subject sphere at any loopback cycle (about every 20-30 milliseconds).

c-3). When the subject reached the corresponding object in the virtual space this object is done and the system generate some special sound to tell it to her/him (say 'bang!'). The real-time loopback interaction ends up when all three subjects reached each correspondent.

Note that two cases are examined about media availability;
Case 1) **Audibility only** : while the loopback runs the user is **put blind on**
Case 2) **Visibility and Audibility :** while the loopback runs the user is provided with the simple 3D wireframe graphics of subjects and objects
The four sound types used for output are 'a:' with higher note when the object is upper front side to the subject, 'a:' with lower note when it is lower front side, 'o:' with higher note when it is upper back side and 'o:' with lower note when it is lower back side.
The average performance of the tests are summarized in the Table 2. The generation and sex of the examinee #1, examinee #2 and examinee #3 are (20's, man), (20's, woman) and (30's, man). Roughly speaking the performance of the reaching objects progresses by repeated practice about several times for all examinees. Comparison between the case 1 and case 2 implies without visibility performance gets worse about ten times. Therefore 3D sound generation technique (e.g, Convolvotron[3]) must be applied to realize more workable system.

5. Conclusions

In this paper we propose the new 3D interaction tool for man-machine interface by using audio media. We show that the application of it is widely spread such as general multiple 3D pointers, amusement tool or, positively, computer aided instructions program for physically handicapped person. This possibility is surely suggested by the demonstration above. In case with visual information, it is needless to say that the ASPECT can support interactive communication more effectively in the field of *virtual reality*.
Our future directions include
1). Parallel manipulation of multiple subjects.
2). Improving the 3D sound generation function with multiple speaker set instead of headphones (which needs a certain echo cancelling technique).
3). Actual applications for the field mentioned above.

References
(1). T.Zimmerman, J.Lanier, C.Blanchard, S.Bryson and Y.Harvill, "A Hand Gesture Interface Device" Proc. of CHI + GI 1987 Conference, ACM, New York, pp.189-192 (1987).
(2). S.S.Fisher, M.McGreevy, J.Humphries and W.Robinett, "Virtual Environment Display System ", Proc.of 1986 Chapel Hill Workshop on Interactive 3D Graphics, Chapel Hill, NC, pp.77-87 (1986).
(3). E.M.Wenzel et.al. "The Convolvotron : Realtime Synthesis of Out-of-Head Localization" Joint Meeting of the Acoustic Soc, of America and Japan (1988).
(4). M.Mitsuishi. et.al. "Tele-Machining System Using Auditory Information" IEICE Technical Report on Human Communications, HC92-19 Japanese pp.51-56 (1992).

Table.1 Correlation among the five vowel sounds

sample vowels		'a'	'i'	'u'	'e'	'o'
test vowels	'a'	0.96	-0.04	-0.14	0.41	0.11
	'i'	-0.02	0.96	0.75	0.02	0.01
	'u'	0.02	0.79	0.92	0.08	0.12
	'e'	0.33	0.02	0.15	0.98	0.22
	'o'	0.16	-0.01	0.13	0.25	0.96

Table.2 Performance results of the experiment

	before practice		after 1 hour practice	
	Case 1	**Case 2**	**Case 1**	**Case 2**
Examinee #1	384sec	203sec	175sec	28sec
Examinee #2	retire	293sec	283sec	41sec
Examinee #3	296sec	82sec	118sec	17sec

Figures

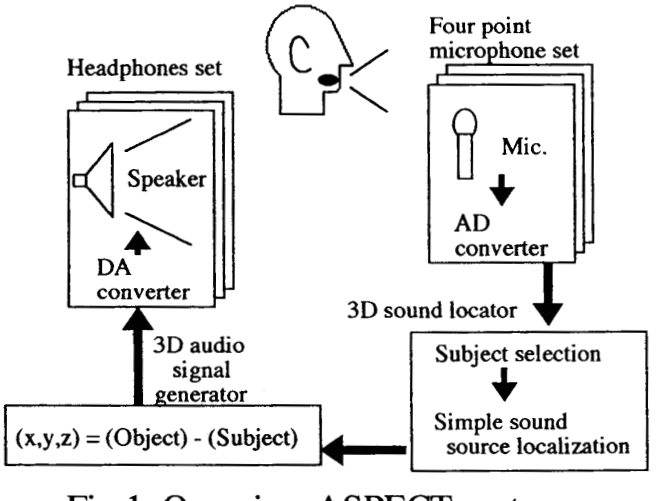

Fig.1 Overview ASPECT system

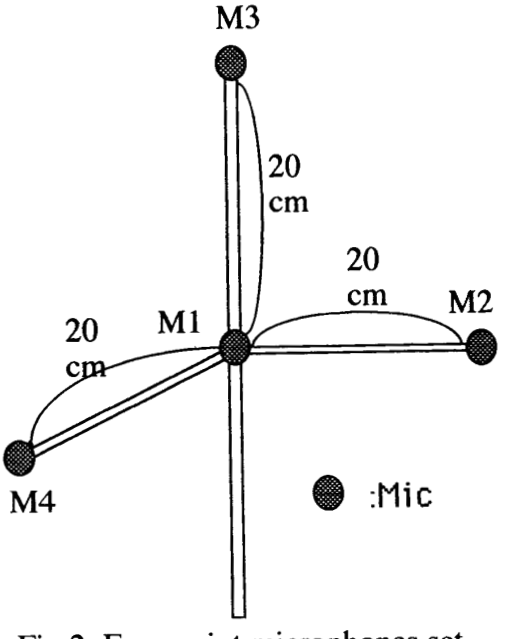

Fig.2 Four point microphones set